A. Experimental details

Here we give more details about how we generate data poisoning attacks under different noise regimes.

Negated loss vs. the original loss.

Definition (Negated loss)**.** Given dataset $(\mathrm{X}, \mathrm{y}) = \{\mathrm{x}_i, y_i\}_{i=1}^n$, the learner returns a function $f(\cdot; \mathrm{w})$. For classification problems, we use the cross entropy loss $L(\mathrm{X}, \mathrm{y}, \mathrm{w}) := \sum_{i=1}^n \ell(y_i, f(\mathrm{x}_i; \mathrm{w}))$, which we also refer as the original loss. We define *the negated loss* as $L'(\mathrm{X}, \mathrm{y}, \mathrm{w}) := \sum_{i=1}^n \ell_{--}(y_i, f(\mathrm{x}_i; \mathrm{w})) = \sum_{i=1}^n -\ell(y_i, -f(\mathrm{x}_i; \mathrm{w}))$.

In particular for binary classification, we use logistic loss $\ell(y_i, f(\mathrm{x}_i; \mathrm{w})) = \log(1+\exp(-y_i f(\mathrm{x}_i; \mathrm{w})))$, whereas the negated loss is $\ell_{--}(y_i, f(\mathrm{x}_i; \mathrm{w})) := -\log(1 + \exp(y_i f(\mathrm{x}_i; \mathrm{w})))$. We plot the relationship between original logistic loss, negated loss and 0-1 loss in Figure 4. Since negated loss is concave, it is a better surrogate loss when we are maximizing the 0-1 loss.

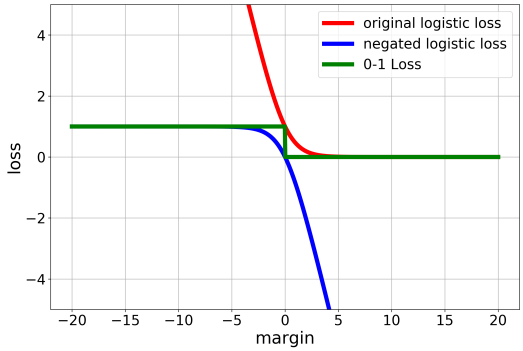

Figure 4. Compare the original loss with the negated loss

Next, we show an additional companion plot to Figure 1 in the main text. Here we generate poisoned data on the proposed negated loss against the original cross entropy loss under ResNet18. The left subplot of Figure 5 shows the clean test accuracy of ResNet18 trained on poisoned CIFAR10 data. The right subplot shows the histogram of L_2 norm of perturbation vectors generated using the two loss functions. As we discuss in the main text (see Section 5 for more details), using the negated loss generates more effective data poisoning attacks.

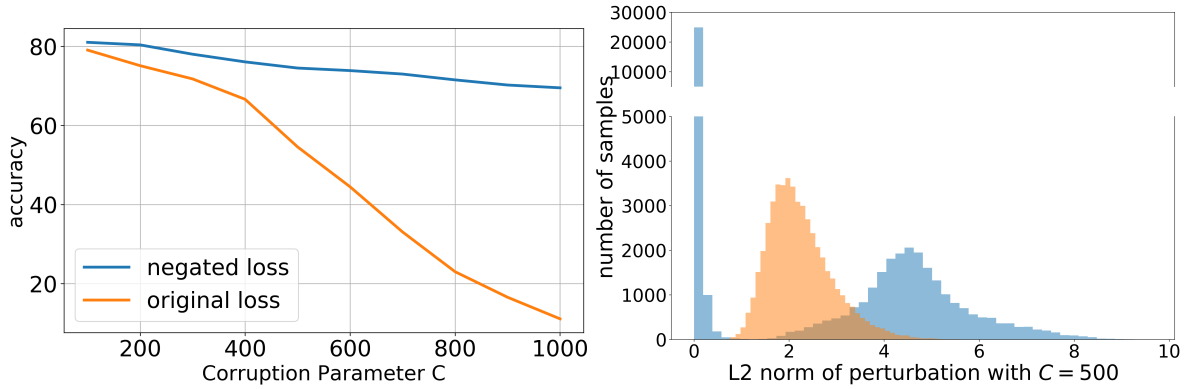

Figure 5. Data poisoning attacks generated by PGD on the proposed negated loss (orange) against the cross entropy loss (blue). The left subplot shows the clean test accuracy of the model trained on poisoned CIFAR10 data with $C = 500$. The right subplot shows the histogram of L_2 norm of perturbation vectors generated using the two loss functions.

Algorithm Details. We use projected gradient ascent on the negated loss to generate the data poisoning attack. The projection operator depends on the noise budget in each data poisoning regime. For noise regime A, we need to guarantee that the set of perturbation vectors satisfy the following norm constraint: $\Delta := \left\{ \{\delta_1, \ldots, \delta_n\} \subset \mathbb{R}^d \middle| \sum_{i=1}^n \|\delta_i\|_2 \leq C\sqrt{n} \right\}$,

where C is the corruption rate, d is the dimension, n is the sample size. For noise regime B, we need to satisfy the following norm constraints: $\Delta := \left\{ \{\delta_1, \ldots, \delta_n\} \subset \mathbb{R}^d \,\middle|\, \|\delta_i\| \leq B, \text{ for } i = 1, \ldots, n \right\}$.

For a detailed description we refer the reader to the pseudocode in Algorithm 1.

Algorithm 1 PGA attack using negated loss

Number of iterations S_1, S_2. Step size $\eta_{\mathrm{w}}, \eta_\delta$. Samples $(\mathrm{X}, \mathrm{y}) = \{\mathrm{x}_i, y_i\}_{i=1}^n$. Perturbation set Δ. Batchsize bs. $\mathcal{P}_\Delta$ is the projection operator onto the set Δ.
Initialize w^0 randomly.
for $s = 0, 1, \ldots, S_1 - 1$ **do**
  for $k = 0, 1, \ldots, \lfloor \frac{n}{bs} \rfloor$ **do**
    Sample a mini-batch of size bs : $\{(\mathrm{x}_{i_1}, y_{i_1}), \cdot, (\mathrm{x}_{i_{bs}}, y_{i_{bs}})\}$. Set $(\hat{\mathrm{x}}_j, \hat{y}_j) = (\mathrm{x}_{i_j}, y_{i_j})$, $j = 1, 2, \cdots, bs$.
    $\mathrm{w}^s = \mathrm{w}^s - \frac{\eta_{\mathrm{w}}}{bs} \sum_{j=1}^{bs} \nabla_{\mathrm{w}} L(\hat{\mathrm{x}}_j, \hat{y}_j, \mathrm{w}^s)$ {mini-batch gradient descent to minimize the original cross entropy loss}
  end for
  $\mathrm{w}^{s+1} = \mathrm{w}^s$
end for
Random initialize: $\delta_i^0 \sim \mathcal{N}(0, \mathrm{I}_d), i = 1, \cdots, n$.
for $s = 0, 1, \ldots, S_2 - 1$ **do**
  for $k = 0, 1, \ldots, \lfloor \frac{n}{bs} \rfloor$ **do**
    Sample a mini-batch of size bs : $\{(\mathrm{x}_{i_1}, y_{i_1}), \cdot, (\mathrm{x}_{i_{bs}}, y_{i_{bs}})\}$. Set $(\hat{\mathrm{x}}_j, \hat{y}_j) = (\mathrm{x}_{i_j}, y_{i_j})$, $j = 1, 2, \cdots, bs$.
    $\delta^s = \delta^s + \frac{\eta_\delta}{bs} \sum_{j=1}^{bs} \nabla_\delta L'(\hat{\mathrm{x}}_j + \delta_j^s, \hat{y}_j, \mathrm{w}^{S_1})$ {Gradient ascent to maximize the negated loss}
    $\delta^s = \min(\max(\mathrm{X} + \delta^s, 0), 1) - \mathrm{X}$ {Clip the pixel value between 0 and 1, then return the perturbation}
  end for
  $\delta^{s+1} = \delta^s$
  $\delta^{s+1} = \mathcal{P}_\Delta(\delta^{s+1})$ {Project onto the set Δ}
end for
return: $\mathrm{X} + \delta^{S_2}$

B. Proofs of theorems

B.1. Proofs of Section 2

We first prove that a bounded perturbation in the input domain corresponds to a bounded perturbation in the gradient domain.

Proof of Proposition 2.1. Recall that for each sample (x_i, y_i) we have $\tilde{\mathrm{x}}_i = \mathrm{x}_i + \delta_i$. Denote the unnormalized margin $\rho := yf(\mathrm{x}; \mathrm{w}), \tilde{\rho} = yf(\tilde{\mathrm{x}}; \mathrm{w})$, where $y = \pm 1$. Then $\ell(yf(\mathrm{x}; \mathrm{w})) = \ell(\rho)$, $\ell(yf(\tilde{\mathrm{x}}; \mathrm{w})) = \ell(\tilde{\rho})$.

$\ell(\rho)$ is L-Lipschitz, i.e. $\|\nabla \ell(\rho)\| \leq L$ for all ρ. $\ell(\rho)$ is α-smooth, i.e. $\|\nabla \ell(\tilde{\rho}) - \nabla \ell(\rho)\| \leq \alpha \|\tilde{\rho} - \rho\|$ for all $\rho, \tilde{\rho}$.

Using the chain rule, we have

$$\begin{aligned}
\|\nabla \ell(yf(\tilde{\mathrm{x}}; \mathrm{w})) - \nabla \ell(yf(\mathrm{x}; \mathrm{w}))\| &= \|\nabla \ell(y\mathrm{w}^\top \mathrm{x}) - \nabla \ell(y\mathrm{w}^\top \tilde{\mathrm{x}})\| \\
&= \|\ell'(\rho) \cdot y\mathrm{x} - \ell'(\tilde{\rho}) \cdot y\tilde{\mathrm{x}}\| \\
&\leq \|\ell'(\rho) \cdot y\mathrm{x} - \ell'(\tilde{\rho}) \cdot y\mathrm{x}\| + \|\ell'(\tilde{\rho}) \cdot y\mathrm{x} - \ell'(\tilde{\rho}) \cdot y\tilde{\mathrm{x}}\| \\
&\leq \|\ell'(\rho) \cdot -\ell'(\tilde{\rho})\| |y| \|\mathrm{x}\| + \|\ell'(\tilde{\rho})\| |y| \|\mathrm{x} - \tilde{\mathrm{x}}\| \\
&\leq \alpha \|y\mathrm{w}^\top \mathrm{x} - y\mathrm{w}^\top \tilde{\mathrm{x}}\| R + L\|\mathrm{x} - \tilde{\mathrm{x}}\| \\
&\leq \alpha |y| \|\mathrm{w}\| \|\mathrm{x} - \mathrm{x}'\| R + L\|\mathrm{x} - \tilde{\mathrm{x}}\| \\
&\leq (\alpha DR + L)\|\mathrm{x} - \tilde{\mathrm{x}}\| \\
&= (\alpha DR + L)\|\delta\|
\end{aligned}$$

□

Now we now prove Theorem 2.2, which shows that SGD is robust against certain bounded adversarial perturbations.

Proof of Theorem 2.2. We define a projection oracle $\mathcal{P}_{\mathcal{W}}$, which given a point $\mathrm{w} \in \mathcal{W}$, returns $\mathcal{P}_{\mathcal{W}}(\mathrm{w}) := \operatorname{argmin}_{\mathrm{v}\in\mathcal{W}} \|\mathrm{w}-\mathrm{v}\|$. Notice that the diameter of $\mathcal{W}$ is $\sup_{\mathrm{w},\mathrm{w}'\in\mathcal{W}} \|\mathrm{w}-\mathrm{w}'\| = \sup_{\mathrm{w},\mathrm{w}'\in\mathcal{W}}\{\|\mathrm{w}\|+\|\mathrm{w}'\|\} = 2D$. Following the standard analysis of SGD for convex functions, we start the analysis by bounding the difference of distances between consecutive iterates w_t and w_{t+1} from w_*:

$$\begin{aligned}
L_{t+1} :&= \mathbb{E}[\|\mathrm{w}_{t+1}-\mathrm{w}_*\|^2] \\
&= \mathbb{E}[\|\Pi_{\mathcal{W}}(\mathrm{w}_t-\eta_t\tilde{\mathrm{g}}(\mathrm{w}_t))-\mathrm{w}_*\|^2] \\
&\le \mathbb{E}[\|\mathrm{w}_t-\eta_t\tilde{\mathrm{g}}(\mathrm{w}_t)-\mathrm{w}_*\|^2] \\
&= \mathbb{E}[\|\mathrm{w}_t-\mathrm{w}_*\|^2] - 2\eta_t\mathbb{E}[\langle\tilde{\mathrm{g}}_t,\mathrm{w}_t-\mathrm{w}_*\rangle] + \eta_t^2\mathbb{E}[\|\tilde{\mathrm{g}}(\mathrm{w}_t)\|^2] \\
&= L_t - 2\eta_t\mathbb{E}[\langle\hat{\mathrm{g}}(\mathrm{w}_t)+\zeta_t,\mathrm{w}_t-\mathrm{w}_*\rangle] + \eta_t^2\mathbb{E}[\|\hat{\mathrm{g}}(\mathrm{w}_t)+\zeta_t\|^2] \\
&\le L_t - 2\eta_t\mathbb{E}[\langle\hat{\mathrm{g}}(\mathrm{w}_t),\mathrm{w}_t-\mathrm{w}_*\rangle] - 2\eta_t\mathbb{E}[\langle\zeta_t,\mathrm{w}_t-\mathrm{w}_*\rangle] + \eta_t^2(\mathbb{E}\|\hat{\mathrm{g}}(\mathrm{w}_t)\|^2 + \|\zeta_t\|^2 + 2\mathbb{E}\langle\hat{\mathrm{g}}(\mathrm{w}_t),\zeta_t\rangle) \\
&\le L_t - 2\eta_t\mathbb{E}[\langle\hat{\mathrm{g}}(\mathrm{w}_t),\mathrm{w}_t-\mathrm{w}_*\rangle] + 2\eta_t\|\zeta_t\|\mathbb{E}[\|\mathrm{w}_t-\mathrm{w}_*\|] + \eta_t^2(G+\|\zeta_t\|)^2 \\
&\le L_t - 2\eta_t\mathbb{E}[\langle\hat{\mathrm{g}}(\mathrm{w}_t),\mathrm{w}_t-\mathrm{w}_*\rangle] + 4\eta_t D\|\zeta_t\| + \eta_t^2(G+\|\zeta_t\|)^2
\end{aligned} \tag{1}$$

Extracting the inner product and averaging over all iterates, we get

$$\begin{aligned}
\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\langle\hat{\mathrm{g}}(\mathrm{w}_t),\mathrm{w}_t-\mathrm{w}_*\rangle] &\le \sum_{t=1}^{T}\frac{L_t-L_{t+1}}{2\eta_t T} + \frac{2D\sum_{t=1}^{T}\|\zeta_t\|}{T} + \sum_{t=1}^{T}\frac{\eta_t(G+\|\zeta_t\|)^2}{2T} \\
&\le \sum_{t=1}^{T}\frac{L_t-L_{t+1}}{2\eta_t T} + \frac{2D\sum_{t=1}^{T}\|\zeta_t\|}{T} + \frac{(G+B')^2}{2T}\sum_{t=1}^{T}\eta_t \\
&\le \frac{L_1-L_{T+1}}{2\frac{D}{\sqrt{T}(G+B')}T} + \frac{2D\sum_{t=1}^{T}\|\zeta_t\|}{T} + \frac{(G+B')^2}{2T}\sum_{t=1}^{T}\frac{D}{\sqrt{T}(G+B')} \\
&\le \frac{4D^2}{2\frac{D}{\sqrt{T}(G+B')}T} + \frac{2D\sum_{t=1}^{T}\|\zeta_t\|}{T} + \frac{D(G+B')}{2\sqrt{T}} \\
&\le \frac{5D(G+B')}{2\sqrt{T}} + \frac{2D\sum_{t=1}^{T}\|\zeta_t\|}{T}
\end{aligned} \tag{2}$$

The result is simply implied by following inequalities:

$$\begin{aligned}
\mathbb{E}[F(\bar{\mathrm{w}})] - F(\mathrm{w}_*) &= \mathbb{E}[F(\frac{1}{T}\sum_{t=1}^{T}\mathrm{w}_t)] - F(\mathrm{w}_*) && \text{(by definition of } \bar{\mathrm{w}}) \\
&\le \mathbb{E}[\frac{1}{T}\sum_{t=1}^{T}F(\mathrm{w}_t)] - F(\mathrm{w}_*) && \text{(by Jensen's inequality)} \\
&= \frac{1}{T}\sum_{t=1}^{T}\big[\mathbb{E}[F(\mathrm{w}_t)] - F(\mathrm{w}_*)\big] && \text{(by linearity of expectation)} \\
&\le \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\langle\hat{\mathrm{g}}(\mathrm{w}_t),\mathrm{w}_t-\mathrm{w}_*\rangle] && \text{(by convexity of } F) \\
&\le \frac{D(G+B')}{\sqrt{T}} + \frac{D\sum_{t=1}^{T}\|\zeta_t\|}{T}
\end{aligned} \tag{3}$$

□

We now prove Theorem 2.3, which shows the rate presented in Theorem 2.2 is optimal in an information-theoretic sense.

Proof of Theorem 2.3. Let $\mathcal{W} = [-1, 1]$ and $F(w) = -wx$, with $x = \begin{cases} +\epsilon & \text{w.p. } 0.5 \\ -\epsilon & \text{w.p. } 0.5 \end{cases}$ for some $\epsilon \leq 1$. Note that the minimum of $F(w)$ is equal to $-|x| = -\epsilon$, which is achieved at $w_* = \operatorname{sgn}(x)$. For simplicity, we denote $\tilde{\mathrm{g}}_t = \tilde{\mathrm{g}}(w_t)$, and $\hat{\mathrm{g}}_t = \hat{\mathrm{g}}(w_t)$. At time t, the oracle returns $\tilde{\mathrm{g}}_t = \hat{\mathrm{g}}_t + \zeta_t$, where $\hat{\mathrm{g}}_t \sim \mathcal{N}(x, 1)$, and $\zeta_t = -\operatorname{sgn}(x)\zeta$ for some $0 \leq \zeta \leq \epsilon$. Therefore, $\tilde{\mathrm{g}}_t \sim \mathcal{N}(x + \zeta_t, 1)$, which is equal to $\tilde{\mathrm{g}}_t \sim \mathcal{N}(\epsilon - \zeta, 1)$ and $\tilde{\mathrm{g}}_t \sim \mathcal{N}(-\epsilon + \zeta, 1)$ if $x = \epsilon$ and $x = -\epsilon$, respectively.

Let $\Psi : \mathbb{R}^T \to [-1, 1]$ be any algorithm that estimates the minimizer of $F(w)$ based on a sequence of T oracle calls. For the simiplicity of presentation, we denote $\Psi_T := \Psi(\tilde{\mathrm{g}}_1, \cdots, \tilde{\mathrm{g}}_T)$. We show that, as long as $T = O(\frac{\sum_{t=1}^T \|\zeta_t\|}{\epsilon} + \frac{1}{\epsilon^2})$, there is a constant probability that Ψ_T fails to identify the sign of x. We first upper bound the probability of success as follows:

$$\begin{aligned}
P(\operatorname{sgn}(\Psi_T) = \operatorname{sgn}(x)) &= \frac{1}{2} P_{\operatorname{sgn}(x)=+1}(\operatorname{sgn}(\Psi_T) = +1)) + \frac{1}{2} P_{\operatorname{sgn}(x)=-1}(\operatorname{sgn}(\Psi_T) = -1) \\
&= \frac{1}{2} - \frac{1}{2} P_{\operatorname{sgn}(x)=+1}(\operatorname{sgn}(\Psi_T) = -1) + \frac{1}{2} P_{\operatorname{sgn}(x)=-1}(\operatorname{sgn}(\Psi_T) = -1) \\
&\leq \frac{1}{2} + \frac{1}{2} TV\left(P_{\operatorname{sgn}(x)=+1}(\tilde{\mathrm{g}}_1, \tilde{\mathrm{g}}_2, \cdots, \tilde{\mathrm{g}}_T), P_{\operatorname{sgn}(x)=-1}(\tilde{\mathrm{g}}_1, \tilde{\mathrm{g}}_2, \cdots, \tilde{\mathrm{g}}_T) \right) \\
&\qquad \text{(Total variation distance definition)} \\
&\leq \frac{1}{2} + \frac{1}{2} \sqrt{\frac{1}{2} D_{KL}\left(P_{\operatorname{sgn}(x)=+1}(\tilde{\mathrm{g}}_1, \tilde{\mathrm{g}}_2, \cdots, \tilde{\mathrm{g}}_T), P_{\operatorname{sgn}(x)=-1}(\tilde{\mathrm{g}}_1, \tilde{\mathrm{g}}_2, \cdots, \tilde{\mathrm{g}}_T) \right)} \\
&\qquad \text{(Pinsker's inequality)}
\end{aligned}$$

Let $\bar{\zeta} = [|\zeta_1|, \cdots, |\zeta_T|] \in [0, 1]^T$ be the vector of (absolute) perturbations. Therefore, we have that $P_{\operatorname{sgn}(x)=+1}(\tilde{\mathrm{g}}_1, \cdots, \tilde{\mathrm{g}}_T) \sim \mathcal{N}(\epsilon \mathbb{1}_T - \bar{\zeta}, \mathrm{I}_T)$, and $P_{\operatorname{sgn}(x)=-1}(\tilde{\mathrm{g}}_1, \cdots, \tilde{\mathrm{g}}_T) \sim \mathcal{N}(-\epsilon \mathbb{1}_T + \bar{\zeta}, \mathrm{I}_T)$. Recall that the KL-divergence between two Gaussian distributions $P_1 = \mathcal{N}(\mu_1, \Sigma_1)$ and $P_2 = \mathcal{N}(\mu_2, \Sigma_2)$ is given by $\mathrm{D}_{KL}(P_1 || P_2) = \frac{1}{2}\left(\operatorname{trace}(\Sigma_2^{-1}\Sigma_1) + (\mu_2 - \mu_1)^\top \Sigma_2^{-1}(\mu_2 - \mu_1) - T + \ln(\frac{\det \Sigma_2}{\det \Sigma_1}) \right)$. Thus, we have that:

$$\begin{aligned}
D_{KL}\left(P_{\operatorname{sgn}(x)=+1}(\tilde{\mathrm{g}}_1, \cdots, \tilde{\mathrm{g}}_T), P_{\operatorname{sgn}(x)=-1}(\tilde{\mathrm{g}}_1, \cdots, \tilde{\mathrm{g}}_T) \right) &= \frac{1}{2} \left(T + \sum_{t=1}^T (2\epsilon - 2|\zeta_t|)^2 - T + \ln(1) \right) \\
&= 2 \sum_{t=1}^T (\epsilon - |\zeta_t|)^2 \\
&\leq 2T\epsilon^2 - 2\epsilon \sum_{t=1}^T |\zeta_t| \qquad (0 \leq |\zeta_t| \leq \epsilon)
\end{aligned}$$

Therefore, we have that $P(\operatorname{sgn}(\Psi_T) \neq \operatorname{sgn}(x)) \geq \frac{1}{2} - \frac{1}{2}\sqrt{T\epsilon^2 - \sum_{t=1}^T \|\zeta_t\|}\epsilon$. Choosing $T = \frac{\sum_{t=1}^T \|\zeta_t\|}{\epsilon} + \frac{1}{4\epsilon^2}$, we have $P(\operatorname{sgn}(\Psi_T) \neq \operatorname{sgn}(x)) \geq \frac{1}{4}$, under which event, the suboptimality gap will at least be $f(\Psi_T) - f(w_*) \geq \epsilon$, and thus

$$\mathbb{E}[F(\Psi_T)] - F(w_*) \geq \frac{1}{4}\epsilon = \frac{\sum_{t=1}^T \|\zeta_t\|}{8T} + \frac{\sqrt{(\sum_{t=1}^T \|\zeta_t\|)^2 + T}}{8T} \geq \frac{3\sum_{t=1}^T \|\zeta_t\|}{16T} + \frac{1}{16\sqrt{T}}.$$

The lower bound holds when the algorithm make at least $O(\frac{\sum_{t=1}^T \|\zeta_t\|}{\epsilon} + \frac{1}{\epsilon^2})$ queries. □

B.2. Proofs of Section 3

We start by recalling the following notations:

$$f_i(\mathrm{W}) = \frac{1}{\sqrt{m}} \sum_{s=1}^m a_s \sigma(\langle \mathrm{w}_s, \mathrm{x}_i \rangle), \tilde{f}_i(\mathrm{W}) = \frac{1}{\sqrt{m}} \sum_{s=1}^m a_s \sigma(\langle \mathrm{w}_s, \tilde{\mathrm{x}}_i \rangle)$$
$$f_i^{(t)}(\mathrm{W}) = \langle \nabla f_i(\mathrm{W}_t), \mathrm{W} \rangle, \tilde{f}_i^{(t)}(\mathrm{W}) = \langle \nabla \tilde{f}_i(\mathrm{W}_t), \mathrm{W} \rangle$$

Also, the following is due to the homogeneity of ReLU:

$$f_i^{(t)}(\mathrm{W}_t) = \langle \nabla f_i(\mathrm{W}_t), \mathrm{W}_t \rangle = f_i(\mathrm{W}_t), \tilde{f}_i^{(t)}(\mathrm{W}_t) = \langle \nabla \tilde{f}_i(\mathrm{W}_t), \mathrm{W}_t \rangle = \tilde{f}_i(\mathrm{W}_t)$$

For any i and W, we define the following quantities:

$$R_i(\mathrm{W}) := \ell(y_i \langle \nabla f_i(\mathrm{W}_i), \mathrm{W} \rangle), Q_i(\mathrm{W}) := -\ell'(y_i \langle \nabla f_i(\mathrm{W}_i), \mathrm{W} \rangle)$$
$$\tilde{R}_i(\mathrm{W}) := \ell(\tilde{y}_i \langle \nabla \tilde{f}_i(\mathrm{W}_i), \mathrm{W} \rangle), \tilde{Q}_i(\mathrm{W}) := -\ell'(\tilde{y}_i \langle \nabla \tilde{f}_i(\mathrm{W}_i), \mathrm{W} \rangle)$$

Again, due to homogeneity of ReLU, we have $R_i(\mathrm{W}_i) = \ell(y_i f_i(\mathrm{W}_i))$, $\tilde{R}_i(\mathrm{W}_i) = \ell(\tilde{y}_i \tilde{f}_i(\mathrm{W}_i))$, and $Q_i(\mathrm{W}_i) = -\ell'(y_i f_i(\mathrm{W}_i))$, $\tilde{Q}_i(\mathrm{W}_i) = -\ell'(\tilde{y}_i \tilde{f}_i(\mathrm{W}_i))$.

Given an initialization $(\mathrm{W}_0, \mathrm{a})$, for any $1 \le s \le m$, define $\bar{\mathrm{u}}_s := \frac{1}{\sqrt{m}} a_s \bar{v}(\mathrm{w}_{s,0})$, where $\bar{v}$ is given by Assumption 1. Collect $\bar{\mathrm{u}}_s$ into a matrix $\bar{\mathrm{U}} \in \mathbb{R}^{m \times d}$. It holds that $\|\bar{\mathrm{u}}_s\|_2 \le \frac{1}{\sqrt{m}}$ and $\|\bar{\mathrm{U}}\|_F \le 1$. Furthermore, note that $\|\nabla f_i(\mathrm{W}_i)\|_F \le 1$, $\|\nabla \tilde{f}_i(\mathrm{W}_i)\|_F \le 1 + \|\delta_i\|_2$.

B.2.1. Proof of Theorem 3.1

We first discuss clean label attacks. Recall that for clean label attack, we have $\tilde{y}_i = y_i$; and for label flip attack, we have $\tilde{f}_i(\mathrm{W}_i) = f_i(\mathrm{W}_i)$. Lemma B.1 ensures that with high probability, the margin attained by $\bar{\mathrm{U}}$ (with respect to features given by the gradient of the network at the initialization) is not much smaller than the margin parameter γ in Assumption 1.

Lemma B.1. [Clean label attacks] Under Assumption 1, given any $\delta \in (0,1)$ and any $\epsilon_1 > 0$, if $m \ge \frac{2\ln(n/\delta)}{\epsilon_1^2}$, then with probability at least $1 - 3\delta$, it holds simultaneously for all $1 \le i \le n$ that

$$y_i \tilde{f}_i^{(0)}(\bar{\mathrm{U}}) = y_i \langle \nabla \tilde{f}_i(\mathrm{W}_0), \bar{\mathrm{U}} \rangle \ge \gamma - \epsilon_1 - C_0 \|\delta_i\|_2 - \frac{\epsilon_1}{2}(1 + B).$$

where $C_0 = 1 + \frac{(4\sqrt{d} + 2\sqrt{\ln(mn/\delta)})}{(1-B)}$.

Proof of lemma B.1. By Assumption 1, given any $1 \le i \le n$,

$$\mu := \mathbb{E}_{\mathrm{w} \sim \mathcal{N}(0, I_d)} \left[y_i \langle \bar{v}(\mathrm{w}), \mathrm{x}_i \rangle \mathbb{1}[\langle \mathrm{w}, \mathrm{x}_i \rangle] > 0 \right] \ge \gamma$$

Let $S'_{i,0} := \left\{ s \middle| \mathbb{1}[\langle \mathrm{w}_{s,0}, \tilde{\mathrm{x}}_i \rangle > 0] - \mathbb{1}[\langle \mathrm{w}_{s,0}, \mathrm{x}_i \rangle > 0] \ne 0, 1 \le s \le m \right\}$ denote the set of hidden nodes at initialization that change sign due to i-th perturbation. Based on Lemma B.3, we know that with probability at least $1 - 2\delta/n$,

$$|S'_{i,0}| \le m \left(\frac{\|\delta_i\|_2}{\|\tilde{\mathrm{x}}_i\|_2} \left(4\sqrt{d} + 2\sqrt{\ln(mn/\delta)} \right) + \frac{\epsilon_1}{2} \|\tilde{\mathrm{x}}_i\|_2 \right)$$

Thus, we have the following set of inequalities:

$$\begin{aligned}
&\left| y_i \tilde{f}_i^{(0)}(\bar{\mathrm{U}}) - y_i f_i^{(0)}(\bar{\mathrm{U}}) \right| = \left| \frac{y_i}{m} \sum_{s=1}^{m} \mathrm{a}_s \left[\langle \bar{v}(\mathrm{w}_{s,0}), \mathrm{x}_i + \delta_i \rangle \mathbb{1}\left[\langle \mathrm{w}_{s,0}, \mathrm{x}_i + \delta_i \rangle > 0 \right] - \langle \bar{v}(\mathrm{w}_{s,0}), \mathrm{x}_i \rangle \mathbb{1}\left[\langle \mathrm{w}_{s,0}, \mathrm{x}_i \rangle > 0 \right] \right] \right| \\
&\le \left| \frac{1}{m} \sum_{s=1}^{m} \mathrm{a}_s \langle \bar{v}(\mathrm{w}_{s,0}), \delta_i \rangle \mathbb{1}\left[\langle \mathrm{w}_{s,0}, \mathrm{x}_i + \delta_i \rangle > 0 \right] \right| + \left| \frac{1}{m} \sum_{s=1}^{m} \mathrm{a}_s \langle \bar{v}(\mathrm{w}_{s,0}), \mathrm{x}_i \rangle \left[\mathbb{1}\left[\langle \mathrm{w}_{s,0}, \mathrm{x}_i + \delta_i \rangle > 0 \right] - \mathbb{1}\left[\langle \mathrm{w}_{s,0}, \mathrm{x}_i \rangle > 0 \right] \right] \right| \\
&\le \|\delta_i\| + \frac{|S'_{i,0}|}{m} \max_s \{ |\langle \bar{v}(\mathrm{w}_{s,0}), \mathrm{x}_i \rangle| \} \\
&\le \|\delta_i\|_2 + \frac{|S'_{i,0}|}{m} \|\mathrm{x}_i\|_2 \\
&\le C_0 \|\delta_i\|_2 + \frac{\epsilon_1}{2}(1 + B)
\end{aligned}$$

where $C_0 = 1 + \frac{(4\sqrt{d}+2\sqrt{\ln(mn/\delta)})}{(1-B)}$. Since $y_i f_i^{(0)}(\bar{\mathrm{U}}) = \frac{1}{m}\sum_{s=1}^{m} y_i \langle \bar{v}(\mathrm{w}_{s,0}), \mathrm{x}_i\rangle \mathbb{1}\big[\langle \mathrm{w}_{s,0}, \mathrm{x}_i\rangle > 0\big]$ is the empirical mean of i.i.d. r.v's supported on $[-1,+1]$ with mean μ, using Hoeffding's inequality, with probability at least $1-\delta/n$,

$$y_i f_i^{(0)}(\bar{\mathrm{U}}) - \gamma \geq y_i f_i^{(0)}(\bar{\mathrm{U}}) - \mu \geq -\sqrt{\frac{2\ln(n/\delta)}{m}} \geq -\epsilon_1. \tag{4}$$

Thus, with probability at least $1-3\delta/n$,

$$\begin{aligned} y_i \tilde{f}_i^{(0)}(\bar{\mathrm{U}}) &\geq y_i f_i^{(0)}(\bar{\mathrm{U}}) - C_0\|\delta_i\|_2 - \frac{\epsilon_1}{2}(1+B) \\ &\geq \gamma - \epsilon_1 - C_0\|\delta_i\|_2 - \frac{\epsilon_1}{2}(1+B). \end{aligned}$$

Applying a union bound finishes the proof. □

The following lemma helps later in the proof of Lemma B.3 to bound the size of the set of hidden nodes whose activation pattern change from initialization.

Lemma B.2. Under Assumption 1, given any $\delta \in (0,1)$ and any $\epsilon_1 > 0$, if $m \geq \frac{2\ln(n/\delta)}{\epsilon_1^2}$, then for any $\epsilon_2 > 0$, with probability at least $1-\delta$, it holds simultaneously for all $1 \leq i \leq n$ that

$$\frac{1}{m}\sum_{s=1}^{m} \mathbb{1}\big[|\langle \mathrm{w}_{s,0}, \tilde{\mathrm{x}}_i\rangle| \leq \epsilon_2\big] \leq \frac{\epsilon_2}{\|\mathrm{x}_i+\delta_i\|_2} + \frac{\epsilon_1}{2}\|\mathrm{x}_i+\delta_i\|_2.$$

Proof of Lemma B.2. Given any fixed ϵ_2 and $1 \leq i \leq n$, we have

$$\mathbb{E}\bigg[\frac{1}{m}\sum_{s=1}^{m} \mathbb{1}\big[|\langle \mathrm{w}_{s,0}, \tilde{\mathrm{x}}_i\rangle| \leq \epsilon_2\big]\bigg] = P\big(|\langle \mathrm{w}, \tilde{\mathrm{x}}_i\rangle| \leq \epsilon_2\big) \leq \frac{2\epsilon_2}{\|\mathrm{x}_i+\delta_i\|_2^2\sqrt{2\pi}} \leq \frac{\epsilon_2}{\|\mathrm{x}_i+\delta_i\|_2},$$

where the expectation is with respect to the randomness in initialization, and the inequality holds since $\langle \mathrm{w}, \tilde{\mathrm{x}}_i\rangle$ is a Gaussian r.v. with variance $\|\mathrm{x}_i+\delta_i\|_2^2$. By Hoeffding inequality, with probability at least $1-\delta/n$,

$$\begin{aligned} &\frac{1}{m}\sum_{s=1}^{m} \mathbb{1}\big[|\langle \mathrm{w}_{s,0}, \tilde{\mathrm{x}}_i\rangle| \leq \epsilon_2\big] \\ &\leq \mathbb{E}\bigg[\frac{1}{m}\sum_{s=1}^{m} \mathbb{1}\big[|\langle \mathrm{w}_{s,0}, \tilde{\mathrm{x}}_i\rangle| \leq \epsilon_2\big]\bigg] + \sqrt{\frac{\ln(n/\delta)}{2m}}\|\mathrm{x}_i+\delta_i\|_2 && \text{(Hoeffding inequality)} \\ &\leq \frac{\epsilon_2}{\|\mathrm{x}_i+\delta_i\|_2} + \frac{\epsilon_1}{2}\|\mathrm{x}_i+\delta_i\|_2 && \big(\epsilon_1 \geq \sqrt{\tfrac{2\ln(n/\delta)}{m}}\big) \end{aligned}$$

Applying a union bound finishes the proof. □

In Lemma B.3, we let $S_{i,t}$ denote the set of neurons at time t that change sign (compared to the initialization) for the i-th poisoned data. We show that the cardinality of $S_{i,t}$ is small as long as the weights stay close to initialization. We also define the set $S'_{i,0}$, which corresponds to the set of neurons that change sign at initialization (time $t=0$) due to the i-th perturbations. We prove that with high probability, it holds that $|S'_{i,0}| \leq \mathcal{O}(m\|\delta_i\|_2)$.

Lemma B.3. For any $1 \leq i \leq n$, denote

$$\begin{aligned} S_{i,t} &:= \big\{s \big| \mathbb{1}[\langle \mathrm{w}_{s,t}, \tilde{\mathrm{x}}_i\rangle > 0] - \mathbb{1}[\langle \mathrm{w}_{s,0}, \tilde{\mathrm{x}}_i\rangle > 0] \neq 0, 1 \leq s \leq m\big\}, \\ S'_{i,0} &:= \big\{s \big| \mathbb{1}[\langle \mathrm{w}_{s,0}, \tilde{\mathrm{x}}_i\rangle > 0] - \mathbb{1}[\langle \mathrm{w}_{s,0}, \mathrm{x}_i\rangle > 0] \neq 0, 1 \leq s \leq m\big\} \end{aligned}$$

If $\|\mathrm{w}_{s,t} - \mathrm{w}_{s,0}\|_2 \leq W_2$, then the following hold:

$$|S_{i,t}| \leq m\bigg(W_2 + \frac{\epsilon_1}{2}\|\tilde{\mathrm{x}}_i\|_2\bigg) \text{ w.p. at least } 1-\delta \text{ for any } 0 \leq \delta \leq 1.$$

$$|S'_{i,0}| \leq m\bigg(\frac{\|\delta_i\|_2}{\|\tilde{\mathrm{x}}_i\|_2}\big(4\sqrt{d} + 2\sqrt{\ln\frac{m}{\delta}}\big) + \frac{\epsilon_1}{2}\|\tilde{\mathrm{x}}_i\|_2\bigg) \text{ w.p. at least } 1-2\delta \text{ for any } 0 \leq \delta \leq 0.5.$$

Proof of Lemma B.3. Note that $s \in S_{i,t}$ implies

$$\big|\langle \mathrm{w}_{s,0}, \tilde{\mathrm{x}}_i\rangle\big| \le \big|\langle \mathrm{w}_{s,0}, \tilde{\mathrm{x}}_i\rangle\big| + \big|\langle \mathrm{w}_{s,t}, \tilde{\mathrm{x}}_i\rangle\big| = \big|\langle \mathrm{w}_{s,t} - \mathrm{w}_{s,0}, \tilde{\mathrm{x}}_i\rangle\big| \le \|\mathrm{w}_{s,t} - \mathrm{w}_{s,0}\|_2 \|\tilde{\mathrm{x}}_i\|_2 \le W_2 \|\tilde{\mathrm{x}}_i\|_2 \tag{5}$$

Let denote $\epsilon_2 = W_2\|\tilde{\mathrm{x}}_i\|_2$. Using Lemma B.2, with probability at least $1-\delta$, we arrive at the following upperbound on the size of $S_{i,t}$:

$$|S_{i,t}| \le \Big|\big\{s \big| |\langle \mathrm{w}_{s,0}, \tilde{\mathrm{x}}_i\rangle| \le \epsilon_2\big\}\Big| = \sum_{s=1}^{m} \mathbb{1}\big[|\langle \mathrm{w}_{s,0}, \tilde{\mathrm{x}}_i\rangle| \le \epsilon_2\big] \le m\bigg(W_2 + \frac{\epsilon_1}{2}\|\tilde{\mathrm{x}}_i\|_2\bigg)$$

Similarly, $s \in S'_{i,0}$ implies

$$\big|\langle \mathrm{w}_{s,0}, \tilde{\mathrm{x}}_i\rangle\big| \le \big|\langle \mathrm{w}_{s,0}, \tilde{\mathrm{x}}_i - \mathrm{x}_i\rangle\big| \le \|\mathrm{w}_{s,0}\|_2 \|\delta_i\|_2$$

Since $\mathrm{w}_{s,0} \sim \mathcal{N}(0, I_d)$, using Gaussian concentration inequality, for any $1 \le s \le m$, with probability at least $1 - \frac{\delta}{m}$, we have that

$$\|\mathrm{w}_{s,0}\|_2 \le 4\sqrt{d} + 2\sqrt{\ln\frac{m}{\delta}}. \tag{6}$$

Let $\epsilon_2 = \big(4\sqrt{d} + 2\sqrt{\ln\frac{m}{\delta}}\big)\|\delta_i\|_2$ and use Lemma B.2. Applying a union bound, we conclude that with probability at least $(1-\delta)(1-\delta) \ge 1 - 2\delta$, the following holds:

$$|S'_{i,0}| \le \Big|\big\{s \big| |\langle \mathrm{w}_{s,0}, \tilde{\mathrm{x}}_i\rangle| \le \epsilon_2\big\}\Big| = \sum_{s=1}^{m} \mathbb{1}\big[|\langle \mathrm{w}_{s,0}, \tilde{\mathrm{x}}_i\rangle| \le \epsilon_2\big] \le m\bigg(\frac{\|\delta_i\|_2}{\|\tilde{\mathrm{x}}_i\|_2}\big(4\sqrt{d} + 2\sqrt{\ln\frac{m}{\delta}}\big) + \frac{\epsilon_1}{2}\|\tilde{\mathrm{x}}_i\|_2\bigg),$$

Which completes the proof. □

Lemma B.4 upperbounds the network output at initialization, under poisoning attacks.

Lemma B.4. Given any $\delta \in (0,1)$, if $m \ge 25\ln(2n/\delta)$, then with probability at least $1-\delta$, it holds simultaneously for all $1 \le i \le n$ that

$$|\tilde{f}_i(\mathrm{W}_0)| = |f(\tilde{\mathrm{x}}_i; \mathrm{W}_0, \mathrm{a})| \le \sqrt{2\ln(4n/\delta)}(1 + \|\delta_i\|_2)$$

Proof of Lemma B.4. The proof closely follows Lemma 2.5 in Ji & Telgarsky (2019). The only difference here is that we bound the norm of $\tilde{\mathrm{x}}_i$ rather than $\|\mathrm{x}_i\|$, which introduces a factor of $1 + \|\delta_i\|$ on the right hand side. Regardless, here, we include a proof for completeness. Given $1 \le i \le n$, let $\mathrm{h}_i = \sigma(\mathrm{W}_0(\mathrm{x}_i + \delta_i))/\sqrt{m} = \sigma(\mathrm{W}_0\frac{(\mathrm{x}_i+\delta_i)}{\|\mathrm{x}_i+\delta_i\|_2})/(\sqrt{m}/\|\mathrm{x}_i + \delta_i\|_2)$. Define $h(a) = \Big(\sum_{i=1}^m \sigma(a_i)^2\Big)^{\frac{1}{2}} = \|\sigma(a)\|_2$, where $\sigma(a)$ is obtained by applying σ coordinate-wise to a. For any $a, b \in \mathbb{R}^m$, by the triangle inequality, we have $|f(a) - f(b)| = |\|\sigma(a)\|_2 - \|\sigma(b)\|_2| \le \|\sigma(a) - \sigma(b)\|_2 = \Big(\sum_{i=1}^m (a_i - b_i)^2\Big)^{\frac{1}{2}} = \|a-b\|_2$. As a result, h is a 1-Lipschitz continuous function w.r.t. the ℓ_2 norm, and $h(\cdot)$ is 1-subgaussian and the bound follows by Gaussian concentration. Thus $\|\mathrm{h}_i\|_2$ is sub-Gaussian with variance proxy $\frac{\|\mathrm{x}_i+\delta_i\|_2^2}{m}$, and with probability at least $1 - \delta/2n$ over W_0,

$$\|\mathrm{h}_i\|_2 - \mathbb{E}[\|\mathrm{h}_i\|_2] \le \sqrt{2\ln\frac{2n}{\delta}}\frac{\|\mathrm{x}_i + \delta_i\|_2}{\sqrt{m}} \le \sqrt{\frac{2}{25}}(1 + \|\delta_i\|_2).$$

On the other hand, by Jensen's inequality,

$$\mathbb{E}[\|\mathrm{h}_i\|_2] \le \sqrt{\mathbb{E}[\|\mathrm{h}_i\|_2^2]} \le \frac{\sqrt{2}}{2}\|\mathrm{x}_i + \delta_i\|_2 \le \frac{\sqrt{2}}{2} + \frac{\sqrt{2}}{2}\|\delta_i\|_2.$$

As a result, with probability at least $1-\delta/2n$, it holds that

$$\|\mathrm{h}_i\|_2 \le \sqrt{\frac{2}{25}}(1+\|\delta_i\|_2)+\frac{\sqrt{2}}{2}+\frac{\sqrt{2}}{2}\|\delta_i\|_2 \le 1+\|\delta_i\|_2.$$

Apply a union bound, with probability at least $1-\delta/2$ over W_0, for all $1\le i\le n$, we have $\|\mathrm{h}_i\|_2 \le 1+\|\delta_i\|_2$. Recall that a is the top layer weights and is uniformly distributed over $\{-1,+1\}$. For any W_0 such that the above event holds, and for any $1\le i\le n$, the r.v. $\langle \mathrm{h}_i, \mathrm{a}\rangle$ is sub-Gaussian with variance proxy $\|\mathrm{h}_i\|_2^2 \le (1+\|\delta_i\|_2)^2$. By Hoeffding's inequality, with probability at least $1-\delta/2n$ over a,

$$|f(\mathrm{x}_i+\delta_i;\mathrm{W}_0,\mathrm{a})| = |\langle \mathrm{h}_i,\mathrm{a}\rangle| \le \sqrt{2\ln(4n/\delta)}(1+\|\delta_i\|_2).$$

With a union bound, with probability $1-\delta/2$ over a, for all $1\le i\le n$, we have $|f(\mathrm{x}_i+\delta_i;\mathrm{W}_0,\mathrm{a})| \le \sqrt{2\ln(4n/\delta)}(1+\|\delta_i\|_2)$. The probability that the above events all happen is at least $(1-\delta/2)(1-\delta/2)\ge 1-\delta$ over W_0 and a. □

Lemma B.5. For any $\overline{\mathrm{W}}$ and any $n\ge 0$, using SGD updates with constant step size $\eta \le \frac{1}{(1+B)^2}$ for $0\le i<n$, then

$$\eta\left(\sum_{i<n}\tilde{R}_i(\mathrm{W}_i)\right)+\|\mathrm{W}_n-\overline{\mathrm{W}}\|_F^2 \le \|\mathrm{W}_0-\overline{\mathrm{W}}\|_F^2+2\eta\left(\sum_{i<n}\tilde{R}_i(\overline{\mathrm{W}})\right)$$

Proof of Lemma B.5. We have

$$\begin{aligned}
\|\mathrm{W}_{i+1}-\overline{\mathrm{W}}\|_F^2 &= \|\mathrm{W}_i-\overline{\mathrm{W}}\|_F^2-2\eta\ell'(\tilde{y}_i\tilde{f}_i(\mathrm{W}_i))\tilde{y}_i\langle\nabla\tilde{f}_i(\mathrm{W}_i),\mathrm{W}_i-\overline{\mathrm{W}}\rangle+\eta^2\left(\ell'(\tilde{y}_i\tilde{f}_i(\mathrm{W}_i))\right)^2\|\nabla\tilde{f}_i(\mathrm{W}_i)\|_F^2\\
&\le \|\mathrm{W}_i-\overline{\mathrm{W}}\|_F^2-2\eta\ell'(\tilde{y}_i\tilde{f}_i(\mathrm{W}_i))(y_if_i(\mathrm{W}_i)-y_if_i^{(i)}(\overline{\mathrm{W}}))+\eta^2(1+B)^2\left(\ell'(\tilde{y}_i\tilde{f}_i(\mathrm{W}_i))\right)^2\\
&\qquad\qquad\text{(Homogeneity of ReLU)}\\
&\le \|\mathrm{W}_i-\overline{\mathrm{W}}\|_F^2-2\eta\tilde{R}_i(\mathrm{W}_i)+2\eta\tilde{R}_i(\overline{\mathrm{W}})+\eta^2(1+B)^2\left(\ell'(\tilde{y}_i\tilde{f}_i(\mathrm{W}_i))\right)^2 \qquad\text{(Convexity of }\ell)\\
&\le \|\mathrm{W}_i-\overline{\mathrm{W}}\|_F^2-2\eta\tilde{R}_i(\mathrm{W}_i)+2\eta\tilde{R}_i(\overline{\mathrm{W}})+\eta\tilde{R}_i(\mathrm{W}_i)\\
&\le \|\mathrm{W}_i-\overline{\mathrm{W}}\|_F^2-\eta\tilde{R}_i(\mathrm{W}_i)+2\eta\tilde{R}_i(\overline{\mathrm{W}})
\end{aligned}$$

where the third inequality is because $\eta^2(1+B)^2\left(\ell'(\tilde{y}_i\tilde{f}_i(\mathrm{W}_i))\right)^2 \le \eta\left(-\ell'(\tilde{y}_i\tilde{f}_i(\mathrm{W}_i))\right)^2 \le -\eta\ell'(\tilde{y}_i\tilde{f}_i(\mathrm{W}_i)) = \eta\tilde{R}_i(\mathrm{W}_i)$. Telescoping gives the result. □

Using Lemma B.1, B.3, B.4, B.5, we prove the following result, which bounds the deviation of iterates from initialization, as well as the average of the instantaneous loss under poisoning attacks.

Lemma B.6. Under Assumption 1, given any risk target $\epsilon\in(0,1)$, and any $\delta\in(0,1/5)$, let

$$\begin{aligned}
\lambda &:= \frac{\ln(4/\epsilon)+\sqrt{2\ln(4n/\delta)}(1+B)}{3\gamma/16-C_0B} > 0,\\
M_1 &:= 4096\lambda^2(1+B)^6/\gamma^6,\\
B_1 &:= \frac{0.04\gamma}{0.04\gamma+4\sqrt{d}+2\sqrt{\ln(mn/\delta)}},
\end{aligned}$$

where $C_0 = 1+\frac{(4\sqrt{d}+2\sqrt{\ln(mn/\delta)})}{(1-B)}$. Then for any $m\ge M_1$, $B\le B_1$, and any constant step size $\eta\le\frac{1}{(1+B)^2}$, with probability at least $1-5\delta$ over the random initialization, if $n := \left\lceil\frac{2\lambda^2}{\eta\epsilon}\right\rceil$, we have 1) for any $0\le i<n$ and any $1\le s\le m$, $\|\mathrm{w}_{s,i}-\mathrm{w}_{s,0}\|_2 \le \frac{4\lambda(1+B)}{\gamma\sqrt{m}}$ holds; 2) $\frac{1}{n}\sum_{i<n}\tilde{R}_i(\mathrm{W}_i)\le\epsilon$.

Proof of Lemma B.6. We follow the proof of Theorem 2.2 in (Ji & Telgarsky, 2019) for data poisoning attack. Let SGD receive $(\tilde{\mathrm{x}}_i, y_i)$ at step i. Let n_1 be the first step before n such that there exists some $1 \le s \le m$ with $\|\mathrm{w}_{s,n_1} - \mathrm{w}_{s,0}\|_2 > \frac{4\lambda(1+B)}{\gamma\sqrt{m}}$. If such a step does not exist, let $n_1 = n$. Let $\overline{\mathrm{W}} := \mathrm{W}_0 + \lambda\bar{\mathrm{U}}$, first show that with probability at least $1 - 5\delta$, for any $0 \le i < n_1$,

$$\tilde{y}_i\langle\nabla\tilde{f}_i(\mathrm{W}_i), \overline{\mathrm{W}}\rangle \ge \ln\left(\frac{4}{\epsilon}\right), \text{ and thus } \tilde{R}_i(\overline{\mathrm{W}}) = \ell(\tilde{y}_i\langle\nabla\tilde{f}_i(\mathrm{W}_i), \overline{\mathrm{W}}\rangle) \le \frac{\epsilon}{4}.$$

Notice that for clean label attack, $\tilde{y}_i = y_i$. We will split the left hand side into three terms and control them separately:

$$y_i\langle\nabla\tilde{f}_i(\mathrm{W}_i), \overline{\mathrm{W}}\rangle = y_i\langle\nabla\tilde{f}_i(\mathrm{W}_0), \mathrm{W}_0\rangle + y_i\langle\nabla\tilde{f}_i(\mathrm{W}_i) - \nabla\tilde{f}_i(\mathrm{W}_0), \mathrm{W}_0\rangle + \lambda y_i\langle\nabla\tilde{f}_i(\mathrm{W}_i), \bar{\mathrm{U}}\rangle \tag{7}$$

- The first term of equation (7) can be controlled using Lemma B.4:

$$|y_i\langle\nabla\tilde{f}_i(\mathrm{W}_0), \mathrm{W}_0\rangle| = |\tilde{f}_i(\mathrm{W}_0)| \le \sqrt{2\ln(4n/\delta)}(1 + \|\delta_i\|_2) \le \sqrt{2\ln(4n/\delta)}(1+B). \tag{8}$$

- The second term of equation (7) can be written as

$$y_i\langle\nabla\tilde{f}_i(\mathrm{W}_i) - \nabla\tilde{f}_i(\mathrm{W}_0), \mathrm{W}_0\rangle = y_i\frac{1}{\sqrt{m}}\sum_{s=1}^{m} a_s\Big(\mathbb{1}[\langle\mathrm{w}_{s,i}, \tilde{\mathrm{x}}_i\rangle > 0] - \mathbb{1}[\langle\mathrm{w}_{s,0}, \tilde{\mathrm{x}}_i\rangle]\Big)\langle\mathrm{w}_{s,0}, \tilde{\mathrm{x}}_i\rangle$$

  Equation (5) and Lemma B.3 ensure that for all $s \in S_{i,i}$, the followings hold:

$$\big|\langle\mathrm{w}_{s,0}, \tilde{\mathrm{x}}_i\rangle\big| \le \big|\langle\mathrm{w}_{s,i} - \mathrm{w}_{s,0}, \tilde{\mathrm{x}}_i\rangle\big| \le \|\mathrm{w}_{s,i} - \mathrm{w}_{s,0}\|_2\|\mathrm{x}_i + \delta_i\|_2 \le \frac{4\lambda(1+B)}{\gamma\sqrt{m}}\|\mathrm{x}_i + \delta_i\|_2$$

$$|S_{i,i}| \le m\left(\frac{4\lambda(1+B)}{\gamma\sqrt{m}} + \frac{\epsilon_1}{2}(1+B)\right)$$

  Let $\epsilon_1 = \frac{\gamma^2}{8(1+B)^3}$. We have,

$$\begin{aligned}\big|y_i\langle\nabla\tilde{f}_i(\mathrm{W}_i) - \nabla\tilde{f}_i(\mathrm{W}_0), \mathrm{W}_0\rangle\big| &\le \frac{1}{\sqrt{m}}\cdot|S_{i,i}|\cdot\big|\langle\mathrm{w}_{s,0}, \tilde{\mathrm{x}}_i\rangle\big| \\ &\le \left(\frac{4\lambda(1+B)}{\gamma\sqrt{m}} + \frac{\epsilon_1}{2}(1+B)\right)\frac{4\lambda}{\gamma}(1+B)\|\mathrm{x}_i + \delta_i\|_2 \\ &\le \frac{\lambda\gamma}{2}\end{aligned} \tag{9}$$

  where the last step is simply plugging into $\epsilon_1 = \frac{\gamma^2}{8(1+B)^3}$ and $m \ge (4096\lambda^2(1+B)^6)/\gamma^6$.

- The third term of equation (7) can be bounded using Lemma B.1,

$$\begin{aligned}y_i\langle\nabla\tilde{f}_i(\mathrm{W}_i), \bar{\mathrm{U}}\rangle &= y_i\langle\nabla\tilde{f}_i(\mathrm{W}_0), \bar{\mathrm{U}}\rangle + y_i\langle\nabla\tilde{f}_i(\mathrm{W}_i) - \nabla\tilde{f}_i(\mathrm{W}_0), \bar{\mathrm{U}}\rangle \\ &\ge \gamma - \epsilon_1 - C_0\|\delta_i\|_2 - \frac{\epsilon_1}{2}(1+B) + y_i\langle\nabla\tilde{f}_i(\mathrm{W}_i) - \nabla\tilde{f}_i(\mathrm{W}_0), \bar{\mathrm{U}}\rangle\end{aligned}$$

  In addition,

$$\begin{aligned}y_i\langle\nabla\tilde{f}_i(\mathrm{W}_i) - \nabla\tilde{f}_i(\mathrm{W}_0), \bar{\mathrm{U}}\rangle &= y_i\frac{1}{m}\sum_{i=1}^{m}\Big(\mathbb{1}\big[\langle\mathrm{w}_{s,i}, \tilde{\mathrm{x}}_i\rangle > 0\big] - \mathbb{1}\big[\langle\mathrm{w}_{s,0}, \tilde{\mathrm{x}}_i\rangle > 0\big]\Big)\langle\bar{v}(\mathrm{w}_{s,0}), \mathrm{x}_i + \delta_i\rangle \\ &\ge -\frac{1}{m}\cdot|S_{i,i}|\cdot\|\mathrm{x}_i + \delta_i\|_2 \\ &\ge -\frac{\gamma^2}{8(1+B)}\end{aligned} \tag{10}$$

Therefore,

$$\begin{aligned} y_i\langle\nabla\tilde{f}_i(\mathrm{W}_i),\bar{\mathrm{U}}\rangle &\geq \gamma - \frac{\gamma^2}{8(1+B)^3} - C_0\|\delta_i\|_2 - \frac{\gamma^2}{16(1+B)^2} - \frac{\gamma^2}{8(1+B)} \\ &\geq \frac{11}{16}\gamma - C_0 B \end{aligned} \tag{11}$$

Put equation (8), (9) and (11) into equation (7), we have

$$\tilde{y}_i\langle\nabla\tilde{f}_i(\mathrm{W}_i),\overline{\mathrm{W}}\rangle = y_i\langle\nabla\tilde{f}_i(\mathrm{W}_i),\overline{\mathrm{W}}\rangle \geq -\sqrt{2\ln(4n/\delta)}(1+B) - \frac{\lambda\gamma}{2} + \lambda(\frac{11}{16}\gamma - C_0 B)$$

Given $\lambda = \frac{\ln(4/\epsilon)+\sqrt{2\ln(4n/\delta)}(1+B)}{3\gamma/16 - C_0 B} > 0$, we have $\tilde{y}_i\langle\nabla\tilde{f}_i(\mathrm{W}_i),\overline{\mathrm{W}}\rangle \geq \ln(\frac{4}{\epsilon})$. Consequently, for any $0 \leq i \leq n_1$, it holds that $\tilde{R}_i(\overline{\mathrm{W}}) \leq \epsilon/4$.

Next we start proving for any $0 \leq i \leq n$ and any $1 \leq s \leq m$, $\|\mathrm{w}_{s,i} - \mathrm{w}_{s,0}\|_2 \leq \frac{4\lambda(1+B)}{\gamma\sqrt{m}}$. Let $n := \left\lceil \frac{2\lambda^2}{\eta\epsilon} \right\rceil$. The next claim is that $n_1 \geq n$. We prove it by contradiction. Suppose $n_1 < n = \left\lceil \frac{2\lambda^2}{\eta\epsilon} \right\rceil$, we start bounding $\|\mathrm{w}_{s,i} - \mathrm{w}_{s,0}\|_2$ using triangle inequality. For any $1 \leq s \leq m$, we have

$$\begin{aligned} \|\mathrm{w}_{s,i} - \mathrm{w}_{s,0}\|_2 &\leq \eta \sum_{\tau<i} \left\| \ell'(\tilde{y}_\tau \tilde{f}_\tau(\mathrm{W}_\tau)) y_\tau \frac{\partial \tilde{f}_\tau}{\partial \mathrm{w}_{s,\tau}} \right\|_2 \\ &\leq \eta \sum_{\tau<i} |\ell'(\tilde{y}_\tau \tilde{f}_\tau(\mathrm{W}_\tau))| \cdot \left\| \frac{\partial \tilde{f}_i}{\partial \mathrm{w}_{s,\tau}} \right\|_2 \\ &\leq \frac{\eta}{\sqrt{m}} \sum_{\tau<i} \tilde{Q}_\tau(\mathrm{W}_\tau)(1 + \|\delta_i\|_2) \end{aligned} \tag{12}$$

Next we start giving the upper bound of $\eta\sum_{i<n_1}\tilde{Q}_i(\mathrm{W}_i)$. To see this, we first use Lemma B.5 to ensure that

$$\|\mathrm{W}_{n_1} - \overline{\mathrm{W}}\|_F^2 \leq \|\mathrm{W}_0 - \overline{\mathrm{W}}\|_F^2 + 2\eta\left(\sum_{i<n_1}\tilde{R}_i(\overline{\mathrm{W}})\right) \leq \lambda^2 + \frac{\epsilon}{2}\eta n_1 \leq 2\lambda^2$$

We further deduct that

$$\sqrt{2}\lambda \geq \|\mathrm{W}_{n_1} - \overline{\mathrm{W}}\|_F \geq \langle \mathrm{W}_{n_1} - \overline{\mathrm{W}}, \bar{\mathrm{U}}\rangle = \langle \mathrm{W}_{n_1} - \mathrm{W}_0, \bar{\mathrm{U}}\rangle - \langle \overline{\mathrm{W}} - \mathrm{W}_0, \bar{\mathrm{U}}\rangle \geq \langle \mathrm{W}_{n_1} - \mathrm{W}_0, \bar{\mathrm{U}}\rangle - \lambda$$

Thus we have $\langle \mathrm{W}_{n_1} - \mathrm{W}_0, \bar{\mathrm{U}}\rangle \leq (\sqrt{2}+1)\lambda$. Moreover, due to equation (11), we arrive at

$$(\sqrt{2}+1)\lambda \geq \langle \mathrm{W}_{n_1} - \mathrm{W}_0, \bar{\mathrm{U}}\rangle = \eta\sum_{i<n_1} -\ell'(y_i\tilde{f}_i(\mathrm{W}_i))y_i\langle\nabla\tilde{f}_i(\mathrm{W}_i),\bar{\mathrm{U}}\rangle \geq \eta\sum_{i<n_1}\tilde{Q}_i(\mathrm{W}_i)(\frac{11}{16}\gamma - C_0 B)$$

Since $B \leq B_1$ gives us $C_0 B \leq 0.08\gamma$, we get that:

$$\eta\sum_{i<n_1}\tilde{Q}_i(\mathrm{W}_i) \leq \frac{(\sqrt{2}+1)\lambda}{11\gamma/16 - C_0 B} \leq \frac{4\lambda}{\gamma}.$$

Plugging the above inequality into (12), we arrive at

$$\|\mathrm{w}_{s,i} - \mathrm{w}_{s,0}\|_2 \leq \frac{4\lambda(1+B)}{\gamma\sqrt{m}}$$

which contradicts the definition of n_1. Therefore $n_1 \geq n$.

Now consider $n := \left\lceil \frac{2\lambda^2}{\eta\epsilon} \right\rceil$. Using Lemma B.5, we can get,

$$\frac{1}{n}\sum_{i<n}\tilde{R}_i(\mathrm{W}_i) \le \frac{\|\mathrm{W}_0-\overline{\mathrm{W}}\|_F^2}{\eta n} + \frac{2}{n}\sum_{i<n}\tilde{R}_i(\overline{\mathrm{W}}) \le \frac{\epsilon}{2}+\frac{\epsilon}{2}=\epsilon. \tag{13}$$

□

Recall that $Q_i(\mathrm{W}) := -\ell'(y_i\langle\nabla f_i(\mathrm{W}_i),\mathrm{W}\rangle)$ is the derivative of the instantaneous loss $R_i(\mathrm{W}) := \ell(y_i\langle\nabla f_i(\mathrm{W}_i),\mathrm{W}\rangle)$. An interesting property of $Q_i(\mathrm{W})$ is that it upperbounds the zero-one loss, and is upperbounded by $R_i(\mathrm{W})$. Lemma B.7 can be proved using a martingale concentration argument.

Lemma B.7 (Lemma 4.3 in Ji & Telgarsky (2019))**.** Define $Q(\mathrm{W}_i) := \mathbb{E}_{(\mathrm{x},y)\sim\mathcal{D}}\big[-\ell'(yf(\mathrm{x};\mathrm{W}_i,\mathrm{a}))\big]$. Given any $\delta\in(0,1)$, with probability at least $1-\delta$, it holds that:

$$\sum_{i\le n}Q(\mathrm{W}_i) \le 4\sum_{i\le n}Q_i(\mathrm{W}_i) + 4\ln\left(\frac{1}{\delta}\right).$$

Using Lemma B.6 and Lemma B.7, we are able to prove our Theorem 3.1 and Theorem 3.2. We argue that under the perturbation budgets considered in our theorems, $R_i(\mathrm{W}_i)$ is closed to $\tilde{R}_i(\mathrm{W}_i)$. In particular, for Theorem 3.1, the crucial step is to show the difference $R_i(\mathrm{W}_i)-\tilde{R}_i(\mathrm{W}_i)$ can be bounded by $\mathcal{O}(\sqrt{md}\|\delta_i\|_2)$ using the convexity of loss function and Lipschitzness of the network.

Theorem B.8 (Regime A, Theorem 3.1)**.** Under Assumption 1, for any $\delta\in(0,1)$, with probability at least $1-\delta$ over random initialization and the training samples, the iterates of SGD with constant step size $\eta=\frac{1}{(1+B)^2\sqrt{n}}$ satisfy

$$\frac{1}{n}\sum_{i<n}L(\mathrm{W}_i) \le \frac{6264(1+B)^2\ln^2(4\sqrt{n})+12528(1+B)^4\ln(24n/\delta)}{\sqrt{n}\gamma^2},$$

provided that $\frac{(1+B)^8\ln(n/\delta)}{\gamma^8}+\frac{(1+B)^6}{\gamma^8}\ln^2(n) \lesssim m \lesssim \frac{(1+B)^4\ln^4(\sqrt{n}/4)+(1+B)^8\ln^2(n/\delta)}{\gamma^4}\frac{n}{S^2}$, and $B \le \frac{0.04\gamma}{0.04\gamma+4\sqrt{d}+2\sqrt{\ln(6mn/\delta)}}$.

Proof of Theorem 3.1. We know that

$$\begin{aligned}
|\tilde{f}_i(\mathrm{W}_i)-f_i(\mathrm{W}_i)| &= \left|\frac{1}{\sqrt{m}}\sum_{s=1}^m a_s\big(\sigma(\langle\mathrm{w}_{s,i},\tilde{\mathrm{x}}_i\rangle)-\sigma(\langle\mathrm{w}_{s,i},\mathrm{x}_i\rangle)\big)\right| \\
&\le \frac{1}{\sqrt{m}}\sum_{s=1}^m\big|\sigma(\langle\mathrm{w}_{s,i},\tilde{\mathrm{x}}_i\rangle)-\sigma(\langle\mathrm{w}_{s,i},\mathrm{x}_i\rangle)\big| \\
&\le \frac{1}{\sqrt{m}}\sum_{s=1}^m\big|\sigma(\langle\mathrm{w}_{s,i},\tilde{\mathrm{x}}_i\rangle-\langle\mathrm{w}_{s,i},\mathrm{x}_i\rangle)\big| && (\sigma \text{ is 1-Lipschitz}) \\
&\le \frac{1}{\sqrt{m}}\sum_{s=1}^m\|\mathrm{w}_{s,i}\|\|\tilde{\mathrm{x}}_i-\mathrm{x}_i\| && (\text{Cauchy-Schwarz}) \\
&= \frac{1}{\sqrt{m}}\sum_{s=1}^m\|\mathrm{w}_{s,i}\|_2\|\delta_i\|_2 \\
&\le \sqrt{m}\|\delta_i\|_2\max_s\|\mathrm{w}_{s,i}\|_2
\end{aligned}$$

where the first inequality is due to Jensen's inequality, and the second inequality holds because the ReLU function is 1-Lipschitz. Since $\mathrm{w}_{s,0}\sim\mathcal{N}(0,I_d)$, we know from (6) that $\|\mathrm{w}_{s,0}\|_2\le 4\sqrt{d}+2\sqrt{\ln(m/\delta)}$ holds with probability at least

$1-\frac{\delta}{m}$ for any $1 \le s \le m$. From Lemma B.6, we have $\|\mathrm{w}_{s,i}-\mathrm{w}_{s,0}\|_2 \le \frac{4\lambda(1+B)}{\gamma\sqrt{m}} \le \frac{\gamma^2}{16(1+B)^2} \le \sqrt{d}$ holds for any $1 \le s \le m$. Combine them together, we arrive at with probability at least $1-\frac{\delta}{m}$,

$$\|\mathrm{w}_{s,i}\|_2 \le \|\mathrm{w}_{s,0}\|_2 + \|\mathrm{w}_{s,i}-\mathrm{w}_{s,0}\|_2 \le (4\sqrt{d}+2\sqrt{\ln(m/\delta)}) + \frac{4\lambda(1+B)}{\gamma\sqrt{m}} \le (5\sqrt{d}+2\sqrt{\ln(m/\delta)}), \tag{14}$$

Take a union bound, with probability at least $1-\delta$, we arrive at

$$\begin{aligned}
\tilde{R}_i(\mathrm{W}_i) - R_i(\mathrm{W}_i) &= \ell(y_i\tilde{f}_i(\mathrm{W}_i)) - \ell(y_i f_i(\mathrm{W}_i)) \\
&\ge \ell'(y_i f_i(\mathrm{W}_i))y_i(\tilde{f}_i(\mathrm{W}_i) - f_i(\mathrm{W}_i)) && \text{(conveixity of } \ell(\cdot)) \\
&\ge -|\tilde{f}_i(\mathrm{W}_i) - f_i(\mathrm{W}_i)| && (-1 \le \ell'(\cdot) \le 0) \\
&\ge -(5\sqrt{d}+2\sqrt{\ln(m/\delta)})\sqrt{m}\|\delta_i\|_2
\end{aligned}$$

Lemma B.6 indicate that $n = \left\lceil \frac{2\lambda^2}{\eta\epsilon} \right\rceil$. Choose $\eta = \frac{1}{(1+B)^2\sqrt{n}}$, we are able to represent ϵ as a function of n:

$$\epsilon \le \frac{2(1+B)^2\lambda^2}{\sqrt{n}} \le \frac{348(1+B)^2\ln^2(4/\epsilon)+696(1+B)^4\ln(4n/\delta)}{\sqrt{n}\gamma^2} \le \frac{348(1+B)^2\ln^2(4\sqrt{n})+696(1+B)^4\ln(4n/\delta)}{\sqrt{n}\gamma^2},$$

where the last inequality follows because $n = \left\lceil \frac{2\lambda^2}{\eta\epsilon} \right\rceil$ implies that $\epsilon \ge \frac{1}{\sqrt{n}}$. Since $m \le \frac{\epsilon^2 n^2}{(5\sqrt{d}+2\sqrt{\ln(m/\delta)})^2 S^2} = \Theta\left(\frac{(1+B)^4\ln^4(4\sqrt{n})+(1+B)^8\ln^2(n/\delta)}{\gamma^4}\frac{n}{S^2}\right)$, combine equation (13), with probability at least $1-5\delta$, we can get

$$\frac{1}{n}\sum_{i<n} Q_i(\mathrm{W}_i) \le \frac{1}{n}\sum_{i<n} R_i(\mathrm{W}_i) \le \frac{1}{n}\sum_{i<n}\tilde{R}_i(\mathrm{W}_i) + (5\sqrt{d}+2\sqrt{\ln(m/\delta)})\frac{S}{n}\sqrt{m} \le \epsilon+\epsilon = 2\epsilon.$$

Further invoking Lemma B.7 gives that with probability at least $1-6\delta$,

$$\frac{1}{n}\sum_{i<n} Q(\mathrm{W}_i) \le \frac{4}{n}\sum_{i<n} Q_i(\mathrm{W}_i) + \frac{4}{n}\ln\left(\frac{1}{\delta}\right) \le 9\epsilon$$

From (Cao & Gu, 2020), we know that $L(\mathrm{W}_i) = P_{(\mathrm{x},y)\sim D}(yf(\mathrm{x};\mathrm{W}_i,\mathrm{a}) \le 0) \le 2Q(\mathrm{W}_i)$. Rescale δ by $1/6$, we have

$$\frac{1}{n}\sum_{i<n} L(\mathrm{W}_i) \le 18\epsilon \le \frac{6264(1+B)^2\ln^2(4\sqrt{n})+12528(1+B)^4\ln(24n/\delta)}{\sqrt{n}\gamma^2}.$$

To make the condition of Lemma B.6 hold, we set the parameters as

$$M_1 = \Omega\left(\frac{(1+B)^8\ln(n/\delta)}{\gamma^8} + \frac{(1+B)^6}{\gamma^8}\ln^2(n)\right).$$

□

B.2.2. Proof of Theorem 3.2

Theorem B.9 (Regime B, Theorem 3.2)**.** Under Assumption 1, for any $\delta \in (0,1)$, with probability at least $1-\delta$ over random initialization and the training samples, the iterates of SGD with constant step size $\eta = \frac{1}{(1+B)^2}$ satisfy

$$\frac{1}{n}\sum_{i<n} L(\mathrm{W}_i) \le \frac{6264(1+B)^2\ln^2(4n)+12528(1+B)^4\ln(24n/\delta)}{n\gamma^2},$$

for $m = \Omega\left(\frac{(1+B)^8\ln(n/\delta)}{\gamma^8} + \frac{(1+B)^6}{\gamma^8}\ln^2(n)\right)$, provided that $B < \min\{\frac{1}{10\sqrt{md}+\sqrt{8m\ln(6m/\delta)}}, \frac{0.04\gamma}{0.04\gamma+4\sqrt{d}+2\sqrt{\ln(mn/\delta)}}\}$.

Now we prove Theorem 3.2. The crucial step here is to show $(1-\mathcal{O}(\sqrt{md}))R_i(\mathrm{W}_i) \le \tilde{R}_i(\mathrm{W}_i)$ using the convexity of the loss function and the fact that $Q_i(\mathrm{W}_i) \le R_i(\mathrm{W}_i)$.

Proof of Theorem 3.2. We have

$$\begin{aligned}
\tilde{R}_i(\mathrm{W}_i) - R_i(\mathrm{W}_i) &= \ell(\tilde{y}_i\tilde{f}_i(\mathrm{W}_i)) - \ell(y_i f_i(\mathrm{W}_i)) \\
&\ge \ell'(y_i f_i(\mathrm{W}_i))y_i(\tilde{f}_i(\mathrm{W}_i) - f_i(\mathrm{W}_i)) && \text{(convexity of } \ell(\cdot)) \\
&\ge -Q_i(\mathrm{W}_i)\max_{1\le i\le n}|\tilde{f}_i(\mathrm{W}_i) - f_i(\mathrm{W}_i)| \\
&\ge -R_i(\mathrm{W}_i)\max_{1\le i\le n}\frac{1}{\sqrt{m}}\sum_{s=1}^{m}\|\mathrm{w}_{s,i}\|_2 B && (Q_i(\mathrm{W}_i) \le R_i(\mathrm{W}_i)) \\
&\ge -\frac{1}{2}R_i(\mathrm{W}_i),
\end{aligned}$$

where the last inequality is because of equation (14) and $B < \frac{1}{2\sqrt{m}(5\sqrt{d}+\sqrt{2\ln(m/\delta)})}$. Thus, $R_t(\mathrm{W}_t) \le 2\tilde{R}_t(\mathrm{W}_t)$. Combine Lemma B.6 equation (13), we can get

$$\frac{1}{n}\sum_{i<n} R_i(\mathrm{W}_i) \le \frac{1}{n}\sum_{i<n} 2\tilde{R}_i(\mathrm{W}_i) \le 2\epsilon.$$

Choose $n = \left\lceil \frac{2\lambda^2}{\eta\epsilon} \right\rceil$, $\eta = \frac{1}{(1+B)^2}$, we are able to represent ϵ as a function of n:

$$\epsilon \le \frac{2(1+B)^2\lambda^2}{n} \le \frac{348(1+B)^2\ln^2(4/\epsilon) + 696(1+B)^4\ln(4n/\delta)}{n} \le \frac{348(1+B)^2\ln^2(4n) + 696(1+B)^4\ln(4n/\delta)}{n},$$

where the last inequality follows because $n = \left\lceil \frac{2\lambda^2}{\eta\epsilon} \right\rceil$ implies that $\epsilon \ge \frac{1}{n}$.

With probability at least $1-5\delta$,

$$\frac{1}{n}\sum_{i<n} Q_i(\mathrm{W}_i) \le \frac{1}{n}\sum_{i<n} R_i(\mathrm{W}_i) \le 2\epsilon.$$

The same procedure as the proof of Theorem 3.1, we get

$$\frac{1}{n}\sum_{i<n} L(\mathrm{W}_i) \le 18\epsilon \le \frac{6264(1+B)^2\ln^2(4n) + 12528(1+B)^4\ln(24n/\delta)}{n\gamma^2}.$$

To make the condition of Lemma B.6 hold, we set the parameters as

$$M_1 = \Omega\left(\frac{(1+B)^8\ln(n/\delta)}{\gamma^8} + \frac{(1+B)^6}{\gamma^8}\ln^2(n)\right).$$

□

B.2.3. Proof of Theorem 3.3

We now focus on label flip attacks. At the i-th iterate, we receive a sample $(\tilde{\mathrm{x}}_i, \tilde{y}_i)$, where $\tilde{\mathrm{x}}_i = \mathrm{x}_i$ and the label is flipped with probability β, that is, $\tilde{y}_i = y_i$ with probability $1-\beta$, and $\tilde{y}_i = -y_i$ otherwise. We first introduce some lemmas that will be used in the proof of the main theorem.

Lemma B.10. Under Assumption 1, given any risk target $\epsilon \in (0,1)$ and any $\delta \in (0,1/3)$, let

$$\lambda := \frac{\sqrt{2\ln(4n/\delta)} + \ln(4/\epsilon)}{\gamma/4},$$
$$M := \frac{4096\lambda^2}{\gamma^6},$$
$$\beta \le \min\{\frac{\epsilon}{12\sqrt{2\ln(4n/\delta)} + \lambda(12+7.5\gamma)}, \frac{(2-\sqrt{3})\epsilon}{3(1+\gamma^2/8)\lambda}\} = \frac{\epsilon\gamma}{(48+42\gamma)\sqrt{2\ln(4n/\delta)} + (48+30\gamma)\ln(4/\epsilon)}$$

Then for any $m \ge M$ and any step size $\eta \le \frac{1}{\sqrt{n}}$, with probability at least $1-3\delta$ over the random initialization, if $n := \lceil \frac{3\lambda^2}{\eta\epsilon} \rceil$, we have 1) $\|\mathrm{w}_{s,i} - \mathrm{w}_{s,0}\|_2 \le \frac{4\lambda}{\gamma\sqrt{m}}$ for any $0 \le i \le n$ and any $1 \le s \le m$; and 2) $\frac{1}{n}\sum_{i<n} \mathbb{E}\ell(\tilde{y}_i \tilde{f}_i(\mathrm{W}_i)) \le \epsilon$. The expectation is with respect to the randomness of label flips.

Proof of Lemma B.10. Let n_1 be the first step before n such that there exists some $1 \le s \le m$ with $\|\mathrm{w}_{s,n_1} - \mathrm{w}_{s,0}\|_2 > \frac{4\lambda}{\gamma\sqrt{m}}$. If such a step does not exist, we simply set $n_1 = n$. As before, let $\bar{\mathrm{W}} := \mathrm{W}_0 + \lambda\bar{\mathrm{U}}$. We first show that with probability at least $1-3\delta$, for any $0 \le i < n_1$, the following holds:

$$y_i\langle \nabla f_i(\mathrm{W}_i), \overline{\mathrm{W}}\rangle \ge \ln\left(\frac{4}{\epsilon}\right)$$

Since in label flip attacks $\tilde{\mathrm{x}}_i = \mathrm{x}_i$, it also holds that $\nabla\tilde{f}_i(\mathrm{W}_i) = \nabla f_i(\mathrm{W}_i)$. We will split the left hand side into three terms and control them individually:

$$y_i\langle \nabla f_i(\mathrm{W}_i), \overline{\mathrm{W}}\rangle = y_i\langle \nabla f_i(\mathrm{W}_0), \mathrm{W}_0\rangle + y_i\langle \nabla f_i(\mathrm{W}_i) - \nabla f_i(\mathrm{W}_0), \mathrm{W}_0\rangle + \lambda y_i\langle \nabla f_i(\mathrm{W}_i), \bar{\mathrm{U}}\rangle. \tag{15}$$

With $\epsilon_1 = \gamma^2/8$, similar to equation (8), (9) and (10) in Lemma B.6 we obtain the following inequalities:

$$\begin{aligned}
&|y_i\langle \nabla f_i(\mathrm{W}_0), \mathrm{W}_0\rangle| \le \sqrt{2\ln(4n/\delta)}.\\
&|y_i\langle \nabla f_i(\mathrm{W}_i) - \nabla f_i(\mathrm{W}_0), \mathrm{W}_0\rangle| \le \frac{\lambda\gamma}{2}.\\
&|y_i\langle \nabla f_i(\mathrm{W}_i) - \nabla f_i(\mathrm{W}_0), \bar{\mathrm{U}}\rangle| \le \frac{\gamma^2}{8}.
\end{aligned} \tag{16}$$

Using inequality (4), with probability at least $1-2\delta$, the third term in Equation (15) can be bounded as follows:

$$\begin{aligned}
y_i\langle \nabla f_i(\mathrm{W}_i), \bar{\mathrm{U}}\rangle &= y_i\langle \nabla f_i(\mathrm{W}_0), \bar{\mathrm{U}}\rangle + y_i\langle \nabla f_i(\mathrm{W}_i) - \nabla f_i(\mathrm{W}_0), \bar{\mathrm{U}}\rangle\\
&\ge (\gamma - \epsilon_1) + y_i\langle \nabla f_i(\mathrm{W}_i) - \nabla f_i(\mathrm{W}_0), \bar{\mathrm{U}}\rangle\\
&\ge \gamma - \gamma^2/4\\
&\ge \frac{3\gamma}{4}.
\end{aligned} \tag{17}$$

Therefore, we get the following lower bound

$$y_i\langle \nabla f_i(\mathrm{W}_i), \overline{\mathrm{W}}\rangle \ge -\sqrt{2\ln(4n/\delta)} - \frac{\lambda\gamma}{2} + \frac{3\lambda\gamma}{4} = \ln(4/\epsilon),$$

Thus, by definition of the logistic loss, for the λ given in the statement of Lemma B.10, we have that:

$$\ell(y_i f_i^{(i)}(\overline{\mathrm{W}})) \le \ln(1+\epsilon/4) \le \epsilon/4.$$

We can similarly give an upper bound on $y_i\langle\nabla f_i(\mathrm{W}_i), \overline{\mathrm{W}}\rangle$, by bounding each term in Equation (15):

$$\begin{aligned} y_i\langle\nabla f_i(\mathrm{W}_i), \overline{\mathrm{W}}\rangle &\le \sqrt{2\ln(4n/\delta)} + \frac{\lambda\gamma}{2} + \lambda y_i\langle\nabla f_i(\mathrm{W}_0), \bar{\mathrm{U}}\rangle + \lambda y_i\langle\nabla f_i(\mathrm{W}_i) - \nabla f_i(\mathrm{W}_0), \bar{\mathrm{U}}\rangle \\ &\le \sqrt{2\ln(4n/\delta)} + \frac{\lambda\gamma}{2} + \lambda|y_i|\|\nabla f_i(\mathrm{W}_0)\|_F\|\bar{\mathrm{U}}\|_F + \frac{\lambda\gamma^2}{8} && \text{(Inequality (16))} \\ &\le \sqrt{2\ln(4n/\delta)} + \lambda(1 + \frac{5\gamma}{8}). \end{aligned}$$

It is easy to see that the logistic loss $\ell(z) = \ln(1+e^{-z})$ satisfies $\ell(-z) - \ell(z) = z$. We leverage this equality to upperbound the instantaneous loss in expectation:

$$\begin{aligned} \mathbb{E}\ell(\tilde{y}_i f_i^{(i)}(\overline{\mathrm{W}})) &= (1-\beta)\ln(1 + e^{-y_i f_i^{(i)}(\overline{\mathrm{W}})}) + \beta\ln(1 + e^{y_i f_i^{(i)}(\overline{\mathrm{W}})}) \\ &= \ell(y_i f_i^{(i)}(\overline{\mathrm{W}})) + \beta y_i\langle\nabla f_i(\mathrm{W}_i), \overline{\mathrm{W}}\rangle \\ &\le \frac{\epsilon}{4} + \beta\left(\sqrt{2\ln(4n/\delta)} + \lambda(1 + \frac{5\gamma}{8})\right) \\ &\le \frac{\epsilon}{4} + \frac{\epsilon}{12} \\ &\le \frac{\epsilon}{3}, \end{aligned} \tag{18}$$

where the penultimate step follows due to the following assumption on the label flip probabilities:

$$\beta \le \frac{\epsilon}{12\sqrt{2\ln(4n/\delta)} + \lambda(12 + 7.5\gamma)}. \tag{19}$$

Let $n := \left\lceil \frac{3\lambda^2}{\eta\epsilon} \right\rceil$; we claim that $n_1 \ge n$. We prove this claim by contradiction. Suppose $n_1 < n$. Using Lemma B.5 with $B = 0$ and taking expectation (with respect to the randomness in label flips) on both side, we have:

$$\mathbb{E}\|\mathrm{W}_{n_1} - \overline{\mathrm{W}}\|_F^2 \le \mathbb{E}\|\mathrm{W}_0 - \overline{\mathrm{W}}\|_F^2 + 2\eta\left(\sum_{i<n_1}\mathbb{E}\ell(\tilde{y}_i f_i(\overline{\mathrm{W}}))\right) \le \lambda^2 + \frac{2}{3}\eta n_1\epsilon \le 3\lambda^2.$$

Further, by Jensen's inequality, we have that $\mathbb{E}\|\mathrm{W}_{n_1} - \overline{\mathrm{W}}\|_F \le \sqrt{\mathbb{E}\|\mathrm{W}_{n_1} - \overline{\mathrm{W}}\|_F^2} \le \sqrt{3}\lambda$. Using $\|\bar{\mathrm{U}}\|_F \le 1$ and the definition of $\overline{\mathrm{W}}$,

$$\sqrt{3}\lambda \ge \mathbb{E}\|\mathrm{W}_{n_1} - \overline{\mathrm{W}}\|_F \ge \langle\mathbb{E}\mathrm{W}_{n_1} - \overline{\mathrm{W}}, \bar{\mathrm{U}}\rangle = \mathbb{E}\langle\mathrm{W}_{n_1} - \mathrm{W}_0, \bar{\mathrm{U}}\rangle - \mathbb{E}\langle\overline{\mathrm{W}} - \mathrm{W}_0, \bar{\mathrm{U}}\rangle \ge \mathbb{E}\langle\mathrm{W}_{n_1} - \mathrm{W}_0, \bar{\mathrm{U}}\rangle - \lambda.$$

The inner product on the right hand side reduces to the following: We have

$$\begin{aligned} \mathbb{E}\langle\mathrm{W}_{n_1} - \mathrm{W}_0, \bar{\mathrm{U}}\rangle &= -\eta\sum_{i<n_1}\langle\mathbb{E}[\ell'(\tilde{y}_i f_i(\mathrm{W}_i))\tilde{y}_i\nabla f_i(\mathrm{W}_i)], \bar{U}\rangle \\ &= \eta\sum_{i<n_1} -(\ell'(y_i f_i(\mathrm{W}_i)) + \beta)y_i\langle\nabla f_i(\mathrm{W}_i), \bar{\mathrm{U}}\rangle \end{aligned} \tag{20}$$

where the last equality is due to the following:

$$\begin{aligned} \mathbb{E}[\ell'(\tilde{y}_i f_i(\mathrm{W}_i))\tilde{y}_i\nabla f_i(\mathrm{W}_i)] &= (1-\beta)y_i\ell'(y_i f_i(\mathrm{W}_i))\nabla f_i(\mathrm{W}_i) - \beta y_i\ell'(-y_i f_i(\mathrm{W}_i))\nabla f_i(\mathrm{W}_i) \\ &= y_i\ell'(y_i f_i(\mathrm{W}_i))\nabla f_i(\mathrm{W}_i) - \beta y_i[\ell'(y_i f_i(\mathrm{W}_i)) + \ell'(-y_i f_i(\mathrm{W}_i)]\nabla f_i(\mathrm{W}_i) \\ &= (\ell'(y_i f_i(\mathrm{W}_i))y_i + \beta)y_i\nabla f_i(\mathrm{W}_i) && (\ell'(-z) + \ell'(z) = -1) \end{aligned}$$

We now rearrange, and lower- and upper-bound the first term in the right hand side of Equation (20):

$$\eta\sum_{i<n_1} -(\ell'(y_i f_i(\mathrm{W}_i)))y_i\langle\nabla f_i(\mathrm{W}_i), \bar{U}\rangle = \mathbb{E}\langle\mathrm{W}_{n_1} - \mathrm{W}_0, \bar{\mathrm{U}}\rangle + \eta\beta\sum_{i<n_1} y_i\langle\nabla f_i(\mathrm{W}_i), \bar{U}\rangle \le (\sqrt{3}+1)\lambda + \eta n_1\beta(1 + \frac{\gamma^2}{8})$$

$$\eta\sum_{i<n_1} -\ell'(y_i f_i(\mathrm{W}_i))y_i\langle\nabla f_i(\mathrm{W}_i), \bar{U}\rangle \ge \eta\sum_{i<n_1} -(\ell'(y_i f_i(\mathrm{W}_i)))\frac{3\gamma}{4} \qquad \text{(Using Equation (17))}$$

where the first inequality follows due to the following:

$$y_i\langle\nabla f_i(\mathrm{W}_i),\bar{U}\rangle = y_i\langle\nabla f_i(\mathrm{W}_0),\bar{U}\rangle + y_i\langle\nabla f_i(\mathrm{W}_i)-\nabla f_i(\mathrm{W}_0),\bar{U}\rangle \le 1+\frac{\gamma^2}{8}.$$

Thus, combining the lower- and the upper-bound above, we get the following bound on the negative sum of the derivative of the instantaneous losses:

$$\eta\sum_{i<n_1}-\ell'(y_i f_i(\mathrm{W}_i)) \le \frac{4}{3\gamma}\left((\sqrt{3}+1)\lambda+\eta n_1\beta(1+\frac{\gamma^2}{8})\right) \le \frac{4\lambda}{\gamma}. \tag{21}$$

where the last inequality holds due to following assumption on the magnitude of the label flip probabilities:

$$\beta \le \frac{(2-\sqrt{3})\epsilon}{3(1+\gamma^2/8)\lambda}.$$

Finally, for any $1\le s\le m$, the distance of the n_1-th iterate from initialization is bounded as:

$$\begin{aligned}
\|\mathrm{w}_{s,n_1}-\mathrm{w}_{s,0}\|_2 &\le \eta\sum_{\tau<n_1}\left\|\ell'(\tilde{y}_\tau f_\tau(\mathrm{W}_\tau))\tilde{y}_\tau\frac{\partial f_\tau}{\partial \mathrm{w}_{s,\tau}}\right\|_2\\
&\le \eta\sum_{\tau<n_1}|\ell'(\tilde{y}_\tau f_\tau(\mathrm{W}_\tau))\tilde{y}_\tau|\cdot\left\|\frac{\partial f_i}{\partial \mathrm{w}_{s,\tau}}\right\|_2\\
&= \frac{\eta}{\sqrt{m}}\sum_{\tau<n_1}-\ell'(\tilde{y}_\tau f_\tau(\mathrm{W}_\tau))\\
&\le \frac{4\lambda}{\gamma\sqrt{m}}. && \text{(Inequality (21).)}
\end{aligned}$$

which contradicts the definition of n_1. Therefore, we conclude that $n_1\ge n$.

Let $n := \left\lceil\frac{3\lambda^2}{\eta\epsilon}\right\rceil$; using Lemma B.5 with $B=0$ and taking expectation on both side, we arrive at

$$\begin{aligned}
\frac{1}{n}\sum_{i<n}\mathbb{E}\ell(\tilde{y}_i\tilde{f}_i(\mathrm{W}_i)) &\le \frac{\|\mathrm{W}_0-\overline{\mathrm{W}}\|_F^2}{\eta n}+\frac{2}{n}\sum_{i<n}\mathbb{E}\ell(\tilde{y}_i f_i^{(i)}(\overline{\mathrm{W}}))\\
&\le \frac{\epsilon}{3}+\frac{2\epsilon}{3}=\epsilon.
\end{aligned} \tag{22}$$

which completes the proof. □

Now we are ready to prove Theorem 3.3.

Theorem B.11 (Regime C, Theorem 3.3)**.** Under Assumption 1, for any $\delta\in(0,1)$, with probability at least $1-\delta$ over random initialization and the training samples, the iterates of SGD with constant step size $\eta=1/\sqrt{n}$ satisfy

$$\frac{1}{n}\sum_{i<n}L(\mathrm{W}_i)\le\frac{1728\ln(12n/\delta)+864\ln^2(\sqrt{n}/4)}{\gamma^2\sqrt{n}},$$

provided $\beta\le\frac{192\ln(12n/\delta)+96\ln^2(\sqrt{n}/4)}{((48+42\gamma)\sqrt{2\ln(16n/\delta)}+(48+30\gamma)\ln(\frac{\gamma^2\sqrt{n}}{48\ln(12n/\delta)+24\ln^2(\sqrt{n}/4)}))\gamma\sqrt{n}}$, and $m=\Omega\left(\frac{\ln(n/\delta)}{\gamma^8}+\frac{1}{\gamma^8}\ln^2(n)\right)$.

Proof of Theorem 3.3. For λ given by Lemma B.10, $n=\left\lceil\frac{3\lambda^2}{\eta\epsilon}\right\rceil$, and $\eta=\frac{1}{\sqrt{n}}$, we have that:

$$\epsilon\le\frac{3\lambda^2}{\sqrt{n}}=\frac{96\ln(4n/\delta)+48\ln^2(\epsilon/4)}{\gamma^2\sqrt{n}}\le\frac{192\ln(4n/\delta)+96\ln^2(\sqrt{n}/4)}{\gamma^2\sqrt{n}}.$$

From Lemma B.10, we need

$$\begin{aligned}\beta &\leq \frac{\epsilon\gamma}{(48+42\gamma)\sqrt{2\ln(4n/\delta)}+(48+30\gamma)\ln(4/\epsilon)} \\ &\leq \frac{192\ln(4n/\delta)+96\ln^2(\sqrt{n}/4)}{((48+42\gamma)\sqrt{2\ln(4n/\delta)}+(48+30\gamma)\ln(\frac{\gamma^2\sqrt{n}}{48\ln(4n/\delta)+24\ln^2(\sqrt{n}/4)}))\gamma\sqrt{n}}.\end{aligned}$$

Recall that $Q(\mathrm{W}_i) := \mathbb{E}_{(\mathrm{x},y)\sim\mathcal{D}}\big[-\ell'(yf(\mathrm{x};\mathrm{W}_i,\mathrm{a}))\big]$. Following (Cao & Gu, 2020), we upperbound the zero-one loss by the negative derivative of the logistic loss, i.e. $L(\mathrm{W}_i) \leq 2Q(\mathrm{W}_i)$. Using Lemma B.10 with δ re-scaled by $1/3$, the following holds with probability at least $1-\delta$,

$$\begin{aligned}\frac{1}{n}\sum_{i<n} L(\mathrm{W}_i) &= \frac{1}{n}\sum_{i<n}\mathbb{P}_{\mathcal{D}}(yf(\mathrm{W}_i)\leq 0) \\ &= \frac{1}{n}\sum_{i<n}\mathbb{P}_{\mathcal{D}}((1-2\beta)yf(\mathrm{W}_i)\leq 0) \\ &\leq \frac{2}{n}\sum_{i<n}\mathbb{E}_{(\mathrm{x},y)\sim\mathcal{D}}\big[-\ell'((1-2\beta)\tilde{y}f(\mathrm{x};\mathrm{W}_i,\mathrm{a}))\big] \\ &\leq \frac{8}{n}\sum_{i<n}-\ell'((1-2\beta)\tilde{y}_i f_i(\mathrm{W}_i))+\epsilon && \text{(Lemma B.7.)} \\ &= \frac{8}{n}\sum_{i<n}-\ell'(\mathbb{E}\tilde{y}_i f_i(\mathrm{W}_i))+\epsilon \\ &\leq \frac{8}{n}\sum_{i<n}\ell(\mathbb{E}\tilde{y}_i f_i(\mathrm{W}_i))+\epsilon && (-\ell'(\cdot)\leq\ell(\cdot)) \\ &\leq \frac{8}{n}\sum_{i<n}\mathbb{E}\ell(\tilde{y}_i f_i(\mathrm{W}_i))+\epsilon && \text{(Jensen's inequality.)} \\ &\leq 9\epsilon && \text{(Lemma B.10)} \\ &\leq \frac{1728\ln(12n/\delta)+864\ln^2(\sqrt{n}/4)}{\gamma^2\sqrt{n}}.\end{aligned}$$

The width requirement in the statement of the theorem comes from plugging the value of λ in the width lower-bound in Lemma B.10. In particular, we have that $m \geq \frac{4096\lambda^2}{\gamma^6}$ and $\lambda = \frac{\sqrt{2\ln(4n/\delta)}+\ln(4/\epsilon)}{\gamma/4}$. Therefore, we get

$$m \geq \frac{4096*32\ln(4n/\delta)+4096*16\ln^2(4/\epsilon)}{\gamma^8}$$

Finally, plugging in $\epsilon \leq \frac{3\lambda^2}{\eta n}$, we get $m = \Omega\bigg(\frac{\ln(n/\delta)}{\gamma^8}+\frac{1}{\gamma^8}\ln^2(n)\bigg)$. □